User's guide to the

Karonga Assessment of Vulnerability Datasets

Rockwool Foundation

December 2012

Table of contents

1 Introduction 4
2 The intervention and its implementation 5
2.1 Products and procedures in groups 6
2.2 Differences between the programme guide and implementation 6
2.2.1 Other implementation issues 7
2.2.2 Variability in treatment intensity 8
3 Research strategy for impact assessment 11
3.1 Timeline of implementation, lottery and data collection 11
3.2 Identifying interested households 11
3.3 The Lottery 12
4 Sample Design and Data Collection 15
4.1 Sampling strategy 15
4.1.1 Sampling lists 15
4.1.2 Sampling Weights 15
4.1.3 Timeline 17
4.2 Actual data collection 17
4.3 Initial non-response 20
4.4 Tracking 21
4.4.1 Rules for tracking 21
4.4.2 Extent of tracking 23
4.5 Attrition 23
4.5.1 Round 2 – 2010 23
4.5.2 Round 3 - 2011 23
4.5.3 Determinants of attrition 24
4.5.4 Split households 24
4.6 Problems encountered in collection of data 24
4.6.1 Round 1 24
4.6.2 Round 2 25
4.6.3 Round 3 25
5 Description of questionnaire modules and variables 26
5.1 Questionnaires 26
5.1.1 Household head questionnaire 26
5.1.2 Female questionnaire 29
5.1.3 Short questionnaire 31
5.1.4 Village questionnaire 32
5.2 Questionnaire modules 33

6 Additional datasets created 33
6.1 All purpose panel cleaned dataset (APPCD) 33
7 Created variables 33
7.1.1 Aggregate consumption 34
7.1.2 Items included in questionnaire 34
7.1.3 Units of measurement 35
7.1.4 Calculating value of consumption in 2009 dataset 36
7.1.5 Calculating calorie intake 2009 38
7.1.6 Calculating value of consumption in the 2010 and 2011 datasets 38
7.1.7 Correlations between the 2009, 2010 and 2011 value of consumption 40
7.1.8 Converting consumption expenditures to USD 41
7.1.9 Suggestions for improvements 41
7.2 Converting R1 agricultural production into kilograms 41
7.3 PAT and PPI 42
7.3.1 PAT 43
7.3.2 PPI 45
8 Using the data 45
8.1 How the raw data file is created 46
8.2 Guide to install and use data 46
8.3 Principles in cleaning the data 47
8.3.1 Which variables to impute/change? 47
8.4 Identification of households 47
8.5 Generating new variables 47
8.6 Creation of panels 48
8.6.1 Questions that are not included in both rounds 48
8.6.2 Sections that are not in both rounds 48
8.6.3 Overview of questionnaire sections and associated panels 48
9 List of abbreviations 52
10 Bibliography 53
11 Annex 1. Schedule of training activities for VSLAs. 57

1 Introduction

This user's guide is made for people who want to use the datasets from Karonga Assessment of Vulnerability. The data was collected in 2009, 2010 and 2011 as part of a cluster randomized control trial. The data might be useful for analyzing other issues than the intervention as it contains detailed background information on a wide range of household characteristics.

This user guide describes how the data was generated. Since the data was collected as a part of an impact assessment, we also briefly describe the intervention and the research strategy for the impact assessment. In general, we have aimed at including aspects relevant to analysis including research strategy, sample design, data collection, questionnaires, and new variables we have created. If you just want to start using the data, you might want to jump to section 8 on using the data. We have included a list of abbreviations in the back of the document. Additional information can be found in the programs (do-files) that cleans the individual data or by asking the team behind the data creation.

From 2014, the data will be publically available and the website of the Rockwool Research Foundation. Until then, any inquiry on data access should be made to Helene BieLilleør.

All data is in a format compatible withStata 11 (.dta).

Good luck with the data!

The research team

Helene BieLilleør
Chris Ksoll
Jonas HelthLønborg
Ole Dahl Rasmussen

2 The intervention and its implementation

The intervention encourages the formation of groups with fifteen to twenty-five members who are trained to manage a village savings and loan association. By and large, it follows the very detailed instructions given in the Programme Guide which can be downloaded from vsla.net(Allen and Staehle, 2007). As no capital is provided, the groups are essentially small financial markets. Members save in and borrow from a common pool of funds kept in a locked cashbox. The inspiration for VSLAs came from rotating savings and credit associations (so-called ROSCAs, see Bouman, 1995), and was developed by CARE international and VSL Associates during the 1990s (Ashe, 2002). The aim has been to improve on ROSCAs in two respects: To make the groups more sustainable as well as more flexible. Sustainability comes from a series of accountability features that prevent theft of funds and elite capture. Flexibility is increased because members can borrow the amount they want, when they want. Whereas ROSCAs multiply without external facilitation, VSLAs only do that to a small degree. This is probably due to the relatively complex procedures designed as safeguards.

The method of setting up village savings and loan associations is thoroughly documented and accompanied by manuals, monitoring tools and frequent workshop on how to implement village savings and loan associations (Allen and Staehle, 2007). The standard process is divided into three phases and is briefly described in the following.

Phase one: Preparation phase (app. 3 weeks). Field officers from the project visits local community leaders as well as local authorities to get permission for visiting the community and to seek support. Through the village headmen, people are invited for information meeting where the project is explained through story telling as described in the manual (Allen and Staehle, 2007). In our case, the first awareness meetings were held for a mixed group of stakeholders consisting Group Village Headmen, Village Headmen, religious leaders, various development committees currently present in the project area as well as government workers at the T/A level (1st quarterly report). Participation ranged from 13 to 129, with average 48 and standard devation 26 (see table below).

Phase two: Intensive phase (app. three months). In the developmental phase, the field officer visit the group at least every week and delivers training sessions on the functioning of the groups. The first five sessions within two weeks, the rest during the remaining two months. Sessions include group leadership and elections, social fund (insurance), buying and selling shares, developing a constitution etc.

Phase three: Maturity and development phase (app 8 months). For the remaining eight months, the field officer supervises the group and makes sure that the VSLA system is followed. Additional training is given as appropriate. Supervision takes place every other week until the group is confident in handling group operations. After this, supervision is bi-monthly. At the end of the period, one training session is delivered on share-out: All accumulated funds in the group are distributed according to the level of savings of each member.

2.1 Products and procedures in groups

Groups work as a member-owned financial intermediary with three products: Savings, credit and insurance. The procedures are described in a constitution. Parts of the constitution is set by a template whereas other parts are decided upon by the participants. The latter include essential aspects like criteria for loans, interest rates and insurance premiums.

Savings are collected at the weekly meetings and conceptualized as buying shares. Each week, a member can buy between one and five shares and the share value is set by the group and written in the group's constitution. Interest on loans is distributed as interest on savings according to the number of shares each member owns. Loans are provided at every fourth meeting and are funded by the savings from members. If the funds requested by members exceed the amount of saved funds, the group decides who gets the loan. Criteria for loans are listed in the constitution and the interest rate on loans is set by the group. Usually, the nominal interest rate on loans is set to between 5 and 20% per month, but extensions in repayment schedules and inflation make the real interest rate considerably lower. Loan contracts runs for three months, with a grace period of one month. The repayment schedule is often adjusted by the group and loans might be repaid over four or five months in need be. The overall interest rate on savings is usually 4 to 5% per month, but materializes only after the end of one cycle when all savings and interest payments are divided by the number of shares and paid out. The difference between the interest rate on loans and interest rate on savings comes from the fact that not all funds are lend out all the time. In the beginning and the end of the cycle, all funds are in the cashbox. The last financial instrument is insurance which is financed by a premium paid by each member each week. The insurance is paid out as a grant or an interest free loan when certain events occur that are outlined in the constitution, usually the death of family members, death of cattle, sudden illness or other emergencies.

Usually after one year of initial training and monitoring, groups “graduate”, which means they are no longer supported by the NGO that helped set them up. Moreover, at that point, there is no government authority supervising them, no savings insurance and no legal documents prove their membership or savings. NGOs can support groups in dealing with issues of theft, but does not normally provide insurance. In our case, we did not provide savings insurance and did not experience theft from groups throughout the period.

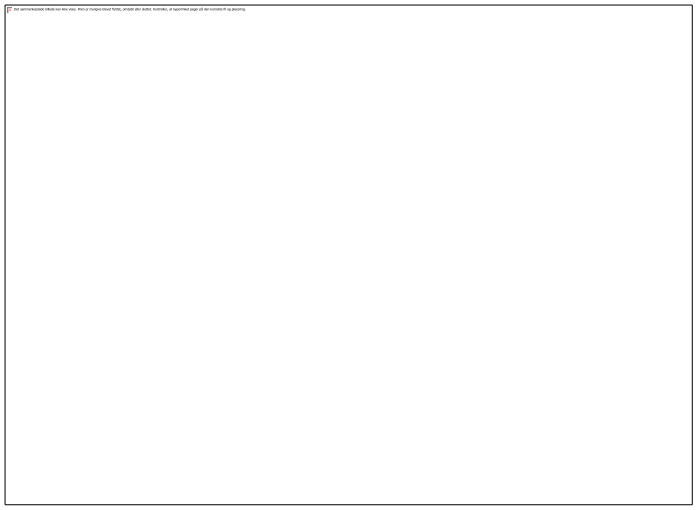
2.2 Differences between the programme guide and implementation

In several instances, the implementation deviated from what is described in the programme guide. Documentation for this comes from quarterly reports, field visits and phone conversations with implementing staff. The central deviations we know of are the following:

- Daily slot savings was not implemented. Daily slot savings is a mechanism that allows for daily savings as described in the programme guide. This system was not implemented in our project, as it was seen as an unnecessary complication by the implementing team.
- The period of training varied between nine and 12 months, i.e. did not necessarily follow the period described in the programme guide. There are two reasons for this: Newer versions of the programme guide work with a nine month

implementation period. And groups were particularly interested in timing share out at the beginning of the agricultural season, thus prolonging the period from beginning savings to share out (see Figure 1).

Figure 1. Cycle length deviates from manual



- Some additional components were added to the methodology to strengthen impact:. These are judged not to pose significant threats to external validity.:
    - Village agents were trained to support implementation. Village agents should start more groups in their own villages and were instructed not to go into control villages. We do not know from the data if a group was started by a project officer or a village agent.
    - Exposure visits. Project officers and village agents visited another VSLA project: Livingstonia Synod Aids Programme in the area of T/A Chikulamayembe in Rumphi district.

2.2.1 Other implementation issues

Apart from the direct deviations as mentioned above, implementation did not go according to plan in a number of instances. Some examples are given below:

- Toward the end of the project, fuel prices sky-rocketed as a result of a general foreign exchange crisis. Due to this, starting groups in control groups was rather slow.
- Implementers indicated that several issues delayed implementation, including lack of time for start-up training due to agricultural work, participation in disbursements of coupons for fertilizer, participation in funerals
- Anecdotal evidence from implementers suggests that rainfall was lower than normal during 2010. This might have affected agricultural output and savings levels in the groups.

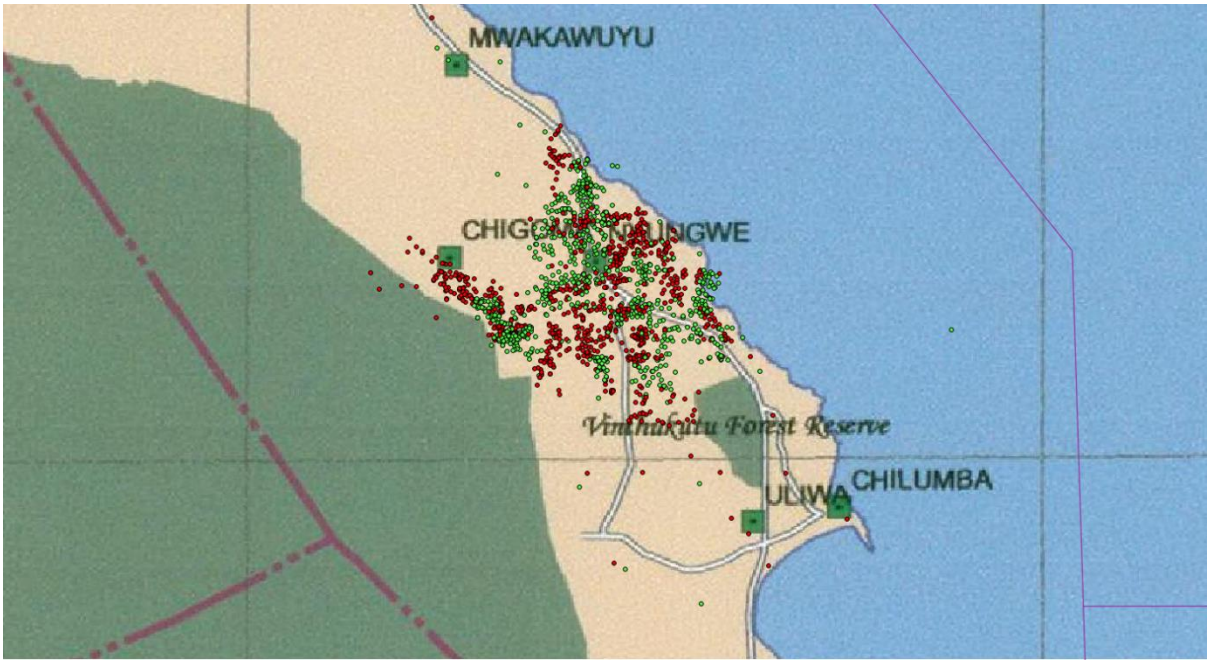
Interest in participation can be divided into three phases: During awareness meetings, many households showed interest. Once savings started, however, many households would initially drop out and entire groups would dismantle. Once each village

had a successful group, however, interest would rise again and household would require more training (e.g. 5th quarterly report).

2.2.2 Variability in treatment intensity

In a typical program like the one being analyzed, whole villages would start at different times. Any effect estimates would therefore be an average of different treatment intensities. In our case, however, almost all villages in the treatment group started at the very beginning. Within each village, however, groups were formed one by one and not all at the same time. This is due to the fact that five field officers had to train all groups and training is particularly intense during the first month. The result is that different people start at different times. The situation is summed up in .

The following figure shows the surveyed households in treatment (green) and control (red) villages at the time of the 2011 survey.



NEEDS INPUT FROM MAP MAKERS:

- Using GPS information to create maps of villages and the surveyed households as well as the timing of the implementation across villages.

Figure 2. Treatment intensity in treatment villages

	Number of groups										Population			Treatment intesity
	2009		2010				2011							
Village name	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**	Village total	Interested in VSLA	Present at awareness meeting	Village pop/number of groups
Bunganiro	1	3	4	4	7	9	9	9	12	13	279	28	41	21
Chakubereka	1	1	1	2	3	6	7	7	7	7	250	26	47	36
Chikalamba		1	1	1	1	1	1	1	1	1	38	14	26	38
Dopa	1	2	2	2	2	2	2	2	2	2	113	34	55	57
Galimoto		1	1	1	1	2	3	3	3	3	110	1	13	37
Kapiyira		2	2	2	2	2	3	3	3	3	193	28	43	64
Mabulungi		2	2	3	4	5	5	5	5	5	157	16	25	31
Makunganya		1	1	1	1	2	2	2	2	2	145	14	21	73
Mbatamila	1	2	2	2	2	2	2	2	2	2	88	15	33	44
Mchekacheka			7	8	9	9	11	11	11	11	469	59	84	43
Mulongoti	1	3	3	3	4	4	4	4	4	4	133	38	89	33
Muyeleka	1	3	3	6	9	9	10	13	15	15	307	48	91	20
Mwachesese			1	2	2	2	2	2	2	2	57	23	31	29
Mwakalomba			1	3	4	4	4	6	7	7	141	31	54	20
Mwakisenjere	2	4	4	4	4	4	4	6	6	6	286	58	93	48
Mwandukutu	1	1	1	1	1	1	1	1	1	1	64	27	41	64
Mwangwala	1	1	1	1	1	1	2	2	2	2	178	37	47	89
Mwayubeyu		2	2	2	3	3	3	3	3	3	119	34	48	40
Mwazolokere 2		1	1	1	1	2	3	3	3	3	174	23	32	58
Mziba	1	1	1	2	2	2	2	2	2	2	76	28	43	38
Ndatira		1	1	3	3	3	3	3	3	3	124	14	28	41
Njalayikwenda	1	1	1	1	2	2	5	5	5	5	100	31	73	20
Nkhubasanga	1	1	1	2	2	2	2	2	2	2	82	17	24	41
Total	*13*	*34*	*44*	*57*	*70*	*79*	*90*	*97*	*103*	*104*	*3683*	*644*	*1082*	

Figure 3. Treatment intensity in control villages

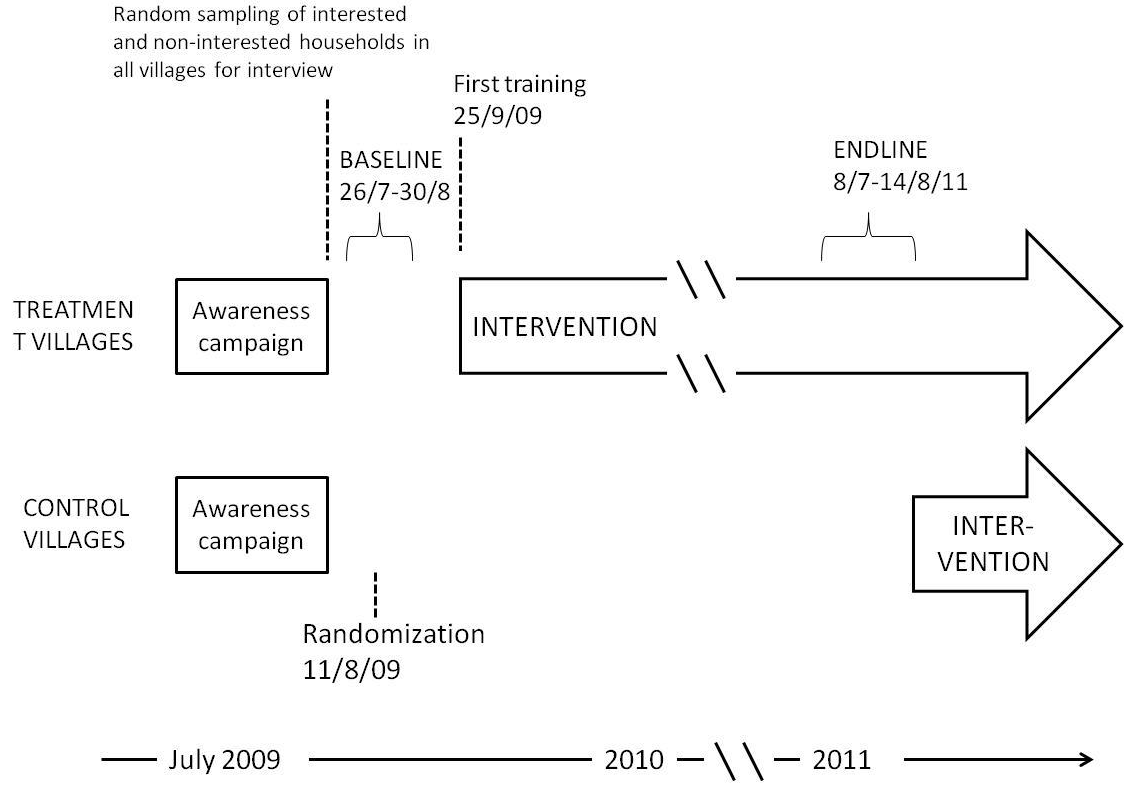
	Number of groups										Population			Treatment intesity
	2009		2010				2011				Village	Interested	Present at awareness	Village pop/number
Village name	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**	total	in VSLA	meeting	of groups
Babalikuni											96	10	24	
Chibwatiko										1	161	14	26	161
Chikombwe										1	175	46	67	175
Chipemphe											78	29	50	
Gangamwale											208	38	64	
Kalimunda											88	20	30	
Kalumwezo											78	15	38	
Kamtembo											80	19	38	
Kanyuka									2	4	190	75	123	48
Kapwambwafu										1	111	29	51	111
Kaswera 1										1	235	19	37	235
Kaswera 2											176	57	82	
Kayaghala											133	16	37	
Masoghaphwire											236	68	129	
Maxwell										1	256	25	43	256
Mkango											58	19	31	
Mmwela											71	0	39	
Mphangweyanjile											178	23	50	
Mudoka											298	15	33	
Mujalebana											129	1	27	
Mwakashunguti											97	32	56	
Mwaseleka											144	14	16	
Mwazolokere											146	18	37	
Total	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*2*	*9*	*3422*	*602*	*1128*	

3 Research strategy for impact assessment

The primary purpose for collection of the data was assessing the impact of the intervention described above. To enable comparison of participating households and the control group, the project was implemented using strategic roll-out (labelled a *randomized order of phase-in* by Duflo, Glennerster and Kremer 2008). Because of this, several aspects of the data collection depends on the research strategy for this particular impact assessment. The strategy, along with the implications for any analysis using the data collected, is described in the following sections.

3.1 Timeline of implementation, lottery and data collection

The implementation of the project, randomization and data collection was highly coordinated which makes an understanding of each element necessary to use the data appropriately. The study employs a two stage stratified sampling procedure. In one stratification, interested households were oversampled purposefully as these were thought to have a higher than average probability of participation. The list of interested households came from lists generated by the Soldev implementation team as will be described in greater detail below.



3.2 Identifying interested households

Before any randomization or data collection, the implementation team created awareness of the upcoming project to the estimated 6.000 households in the 46 villages of Mwirang'ombe in the Karonga district. This was done by Soldev field officers who first went to each village to meet with the village headman. The village headman was briefly introduced to the concept, and asked to gather the villagers at a later point in time, at which the field officer would return. At this second visit the Soldev field officer would introduce the concept thoroughly to the attendees and finally

asked them to indicate whether they were interested in participation. Villagers were informed that prior to the start-up of the project and understanding of the extent of interest across villages was necessary to achieve the best possible implementation of program. They were furthermore informed that the project will be initiated in their village at some point within the coming three years.

At the end of the awareness meetings any individuals interested in participating were asked to sign up. Interested individuals were asked to record their names, as well as the name of the household head and the spouse of the household head from the household in which they resided. These lists were subsequently collected by Soldev staff and merged to census lists of households in the TA Mwirangombe for survey sample selection. The census lists were provided by the Karonga Agricultural Development Division (ADD) who use them for distributing coupons for the governmental fertilizer and grain subsidy (see e.g. Dorward&Chirwa 2011).

A number of problems occurred in the process of matching the names of the household heads from our lists of interested households with the census lists. Some matches hinged on a subjective assessment of the match as names were spelled differently (e.g. due to the interchangeable use of the letters 'l' and 'r' in Chichewa. If the differences were minor, the match was accepted as a unique match.

3.3 The Lottery

Creating treatment and control groups that are as homogenous as possible is of cause crucial in any randomized controlled trial. With an increasing number of units available for randomization the higher the likelihood of obtaining a balanced treatment and control group. We employed blocked randomization to obtain better balance.

The blocks were decided upon by the researchers in collaboration with Soldev field officers to ensure conformability with the research strategy. This also ensured utilization of local knowledge regarding the otherwise unobservable characteristics of the villages. The villages where divided into 7 groups: large fishing villages, other fishing villages, large non-fishing villages, eager villages, villages around Wovwe, villages were World Vision had previously initiated a similar intervention and a group of remaining villages. Some definitions are in order:

- Large villages (4 fishing and 4 non-fishing villages):
  The four largest villages with respect to number of households as suggested by the ADD census lists were chosen for the large villages-category. This meant that the large fishing villages had between 234 and 303 households and the large non-fishing villages had between 177 and 326 households.
- Fishing villages (8 villages):
  Villages within 1 kilometre of the lake were defined as fishing villages. The 8 fishing villages do not include the 4 largest fishing villages i.e. 12 villages were identified as fishing villages in total.
- Eager villages (7 villages):
  These were defined by the Soldev field officers after having done the awareness meetings in the villages. They identified the 7 villages where the villagers seemed most eager to start the project. In several cases this meant they had already formed groups and were ready to start training even before the field officers had asked them to.

- World Vision Villages (10 villages):
  World Vision had initiated a similar savings based project in the villages of one group village headman (GVH Mwakasyunguti) a couple of years earlier. Even though the project had not been that successful, and no groups to our knowledge were still functioning the villages were still grouped into one for the lottery, assuring half the villages would be in the treatment and half in the control group.
- Villages around Wowve (2 villages)
  This was primarily done for making training possible for one of the Synodev staff who were pregnant at the beginning of the project and would thus be unable to travel long distances by motorbike which would be required to reach the other villages in the area.
- Remaining villages (11 villages):
  Villages not falling into any of the above categories

Having assigned each of the villages into the above blocks, the lottery itself could be undertaken. The lottery was undertaken by the core researches along with the four Soldev field officers and a representative from the Soldev head office on August 11th at the house of the Reverend Mchulu in Nyungwe. The lottery was conducted by putting all village names from a specific group written into a bowl and drawing half the names. In the blocks where there was an uneven number of villages in the group (the group of eager villages and the group of remaining villages) a pre-lottery was conducted to assess which of the groups would have an extra village chosen as a treatment village. The outcome of the lottery was as follows:

Table 1: Outcome of lottery

Village name	Treatment	Control
Large Fishing		
Muyeleka	X	
Mudoka		X
Mwakasenjere	X	
Masoghafwile		X
Other fishing		
Chibwatiko		X
Kaswera I		X
Kaswera II		X
Kayaghala		X
Makonganya	X	
Mlongoti	X	
Ndatila	X	
Galimoto	X	
Eager		
Mujalebana		X
Mwazolokere II	X	
Mwela		X
Kapwambafu		X
Mwayubeya	X	
Mbatamira	X	
Mwakalomba	X	
Wowve		
Kanyuka		X
Mwangwala	X	
Large non-fishing		
Gangamwale		X
MchekaCheka	X	
Mphangeweyanjini		X
Bunganilo	X	
World Vision		
Chiphemphe		X
Dopa	X	
Kalumwezo		X
Makutembo		X
Mukango		X
Mwachesese	X	
Mwakasyunguti		X
Mwandukutu	X	
Mziba	X	
Nkhutabasanga	X	
Remaining		
Babalukuni		X
Mabulungi	X	
Kapiyira	X	
Kalimunda		X
Maxwell		X
Chikombwe		X
Chikalamba	X	
Njalayikwenda	X	
Mwazolokele I		X
Mwaseleka		X
Chakubeleka	X	

4 Sample Design and Data Collection

4.1 Sampling strategy

Once the matching of the lists of interested households with the ADD census was completed for the villages, the survey sampling could be initiated. A two-stage stratified sampling was employed. The first stage involved sampling the same absolute number of households from each village. This was done irrespective of the village size as the level of randomization was at the village level, and hence desire was for minimizing the variance in the village level outcomes such as consumption. Secondly, each village was divided into interested and non-interested households. In each of the two groups, each observation was assigned a random number between 0 and 1. The households were then sorted in descending order. From the list of interested households, the first 14 observations were assigned to the long questionnaire and the following ten to the short questionnaire. Then the next four were used as replacements for the long questionnaire and the remainder as replacement for the interested short questionnaire interviews. A similar strategy was imposed for the non-interested households, except in this case it was only the first four observations that were subject to a long questionnaire, ten for the short questionnaire, two replacement households for the long questionnaire and a number of replacement households for the short questionnaire.

Table 2: Distribution of questionnaires within each village

	Interested households	**Non-interested households**
Long questionnaire	14	4
Short questionnaire	10	10

With a total of 46 villages sampled, this should give a total sample of 828 long questionnaires and 920 short questionnaires.

4.1.1 Sampling lists

…

4.1.2 Sampling Weights

Due to the oversampling of interested households sampling weights must be used in any analysis generating population-wide results. In general we follow UN (2005) below.

Two weight variables are created for all datasets. Which weight to use depends on the dataset used. If only the long questionnaire data is used, one weight must be applied. If both the long and short questionnaire datasets are used, another weight must be used in the analysis.

The sampling weights used are the inverse probability of being sampled. We stratify on two variables: Village and interest. Since we have 46 village, but two villages had no interested households, this result in 90 strata.

The inverse probability of being sampled is simply the proportion of observations in the data within a stratum relative to the population in the same stratum.

4.1.2.1 Intuition

A useful interpretation of the use of the inverse probability of sampling (for now call it the *ips)* as sample weights is that *ips* should be the number of households an observation in the sample represents. The intuition is simple enough: if we sample 50 out of 100 households, each observation represents two households and ips=2. Now, if we have a response rate of 50%, the observations do not represent two households each, but four households each. Formally, we can adjust the weight with the inverse response rate: If the 'base weight' is two and the response rate is 0.5, then the actual weight should be 2*1/0.5=4. In our calculation, we simply just use the number in the data, which is already adjusted for response rate.

Following the comment in the introduction, we generate two different sets of sampling weights: one set for analysis using only the long questionnaire data and another set for analysis using both short and long questionnaire data.

The subscript denotes whether the weight is for the interested (I) or non-interested (NI) households, and whether the weight applies to the long questionnaire data (L) or the combination of short and long questionnaire data (LS). SI_v and SN_v is the number of sampled interested in the village and total number of sampled in the village, respectively. I_v and N_v denotes the number of interested and total number of households in village v respectively.

4.1.2.2 Using data from long and short questionnaires

Sampling weight for interested households in village v

$$W^v_{I,LS} = \frac{I_v}{SI_v}$$

Sampling weight for non-interested households

$$W^v_{NI,LS} = \frac{NI_v}{SN_v - SI_v}$$

In the data this weight is called *allweight*.

4.1.2.3 Using long questionnaire data only

All the observations from the long questionnaires can be regarded as a smaller, but otherwise similar, random sample representing the same population. As such, the nominator is the same.

Sampling weight for interested household

$$W^v_{I,L} = \frac{NI_{lv}}{SI_{lv}}$$

Sampling weight for non-interested household

$$W_{NI,L}^{v} = \frac{NI_{lv}}{SN_{lv} - SI_{lv}}$$

This weight is called *weightlong* in the data.

4.1.2.4 Using only the interested households

For analysis on the interested households only, the weight *allweight* should be used. The reason is the simple that the population is reduced and the number of people in population and sample for the interested households remain unchanged.

4.1.2.5 Non-response

Non-response is a legitimate concern. During surveying, non-responding households were replaced with other, randomly drawn households. The non-responding households might be systematically different from the responding and the replacements which would create a bias compared to the population.

With our data it is not possible to carry out post stratification to make the sample representative, but an analysis of the severity of the problem is made in section 4.3.

4.1.3 Timeline

Data was collected after the awareness meetings and the randomization of villages into treatment and control groups was completed as described in section 3.1. The first round of data was collected from July 26 2009 to August 30 2009. The second round of data collection occurred between August 18 and September 20 2010. The final round of data was collected between July 8 and August 14 2011.

4.2 Actual data collection

Data was collected by The Invest in Knowledge Initiative (IKI) in all three rounds. Among other advantages this also ensured consistency in the individuals visiting the surveyed households. To as wide extent as possible interviewers were sent back to the households they had interviewed in the 2009 survey to ensure a high level of trust between interviewees and interviewers.

In the 2009 and 2010 surveys the interviews were undertaken by three teams. Each team consisted of a supervisor, eight interviewers and a driver. Of the eight interviewers, six were normally allocated to undertaking the long questionnaires and the remaining two to do the short questionnaires. For the 2009 and 2010 surveys, these two interviewers were provided with a PDA on which the short questionnaire had been programmed. Paper and pencil was used by the short questionnaire interviewers for writing the names of the household members. In the instances where a long questionnaire interviewer did not have time to complete an additional head or female interview on a given day, they were assigned to complete a short interview completed on paper.

In the 2011 survey four teams were used. Each team consisted of a supervisor, six interviewers and a driver. The reason for the smaller teams was the increased focus on tracking households both within and outside the survey area. In the 2011 survey all short questionnaire interviews were conducted on paper. An interviewer was expected to complete three long questionnaires (1½ household) on a given work day or

eight short questionnaires. The long questionnaires took approximately two hours to complete each (i.e. a household were interviewed for a total of four hours) and the short questionnaires approximately 45 minutes.

For the 2009 survey the supervisor was given the list of survey respondents, and the responsibility of completing the villages. A village was considered complete when 14 interested and 4 non-interested households had been subjected to the long questionnaire, a total of 20 short questionnaires (10 interested and 10 non-interested where possible) had been undertaken and a village questionnaire was completed.

In the 2010 and 2011 survey the supervisors were handed prefilled questionnaires, i.e. the questionnaires already contained information on time-invariant variables, such as the gender and age of existing household members as well as lists of the activities the households were engaged in the year before. This was done both to save time in the interview, but more importantly to ensure information was collected about the current state of the activities and members that comprised the householdat the time of the previous survey. The researchers associated with the project cleaned the first rounds of data and subsequently added the information about the names, age and gender of the household members, the enterprises the household was engaged in as well as the number of livestock owned at the time of the previous survey in the questionnaires used for the subsequent rounds of data collection. As an added bonus this also verified the information recorded in previous surveys.

The final sample consists of 833 long questionnaires and 942 short questionnaires from the 46 villages, distributed in the following way:

Table 3: Actual distribution of questionnaires across villages, 2009

	Number of completed interviews					
	Long Questionnaire			Short Questionnaire		
Village	I=1	I=0	Total	I=1	I=0	Total
Babalikuni	7	11	18	0	21	21
Bunganiro	14	4	18	7	16	23
Chakubereka	14	4	18	4	17	21
Chibwatiko	9	9	18	0	21	21
Chikalamba	11	6	17	0	3	3
Chikombwe	17	1	18	13	8	21
Chipemphe	7	11	18	10	9	19
Dopa	9	9	18	7	18	25
Galimoto	0	18	18	0	18	18
Gangamwale	15	3	18	9	12	21
Kalimunda	15	3	18	2	19	21
Kalumwezo	6	12	18	2	5	7
Kamtembo	14	4	18	3	20	23
Kanyuka	14	4	18	7	14	21
Kapiyira	15	3	18	6	17	23
Kapwambwafu	13	5	18	9	10	19
Kaswera 1	15	4	19	0	21	21
Kaswera 2	14	5	19	10	15	25
Kayaghala	12	6	18	1	20	21
Mabulungi	10	8	18	0	21	21

Makunganya	14	4	18	0	22	22
Masoghaphwire	15	4	19	10	14	24
Maxwell	14	4	18	5	16	21
Mbatamila	14	4	18	0	20	20
Mchekacheka	15	3	18	8	13	21
Mkango	14	4	18	0	20	20
Mmwela	0	17	17	0	7	7
Mphangweyanjile	14	4	18	3	18	21
Mudoka	14	4	18	0	22	22
Mujalebana	1	17	18	0	21	21
Mulongoti	14	4	18	8	14	22
Muyeleka	14	4	18	9	14	23
Mwachesese	13	5	18	6	12	18
Mwakalomba	14	5	19	7	15	22
Mwakashunguti	11	8	19	10	13	23
Mwakisenjere	15	5	20	9	12	21
Mwandukutu	14	4	18	8	13	21
Mwangwala	15	3	18	9	15	24
Mwaseleka	11	7	18	0	21	21
Mwayubeyu	14	4	18	7	13	20
Mwazolokere	10	8	18	0	18	18
Mwazolokere 2	13	5	18	3	18	21
Mziba	9	9	18	5	17	22
Ndatira	12	6	18	0	23	23
Njalayikwenda	14	4	18	7	14	21
Nkhubasanga	13	5	18	1	27	28
Total	552	281	833	205	737	942

As is evident from the table above, the actual distribution of surveyed households does not follow the planned distribution given in table 2. There are several explanations for this. The first is that there might not have been 24 interested households in each village. In the village of Mmwela, for instance, no interested households were located, and hence all surveyed households were sampled from the census lists provided by ADD.

The lack of interested households cannot, however, explain the fact that there were only surveyed 11 interested households for the long questionnaire in the village of-Mwakashunguti. Here another 10 interested households were administered the short questionnaire. Even though supervisors were instructed to give priority to the completion of 18 (14+4) long questionnaire, practically the long and short questionnaires were administered simultaneously. If a number of interested households scheduled for the long questionnaire could not be located or refused participation, there might not have been a sufficient number of replacement households. If, at the same time, the interested households scheduled for the short questionnaire had already been surveyed, they were not re-administered the long questionnaire.

In the case were the interested person was the household head, and this had no spouse, the interested individual was subject to both the “Female” and “Head” ques-

tionnaires. Care was taken not to undertake the two interviews back to back, but to arrange the two interviews at two different days to prevent fatigue.

4.3 Initial non-response

Households were only replaced in the 2009 baseline data collection. In the 2010 and 2011 data collection rounds, households that were interviewed in the 2009 survey were re-visited. Household that could not be revisited at this time were not replaced by other households (see the section on attrition below).

As described above, the supervisor was allowed to replace households selected for interviews in the following four cases:

a. When the designated respondent or the spouse of the household head was above 65 years of age
b. When the household could not be identified (i.e. had moved, or the household head had passed away and no other person had overtaken the household)
c. When the designated respondent or the spouse of the household head was away for the total period in which the interviewers where in the area, and no rescheduling of the interview could be agreed upon
d. When the respondents refused to participate in the interview even after careful explanation of the contributions of his/her time to the research conducted.

Although non-response was not considered a major problem by the survey team, a number of replacement households had to be used in order to gather the required number of observations from each village.

The following tables show the number of actually surveyed households and their status according to the sampling for the long and the short questionnaire respectively.

Table 4: Type of sampling for surveyed households

Type of sampling, surveyed long Q respondents

	Freq.	Percent	Cum.
Long Q, interested	458	54.98	54.98
Long Q, non-interested	113	13.57	68.55
Short Q, interested	11	1.32	69.87
Short Q, non-interested	25	3	72.87
Replacement, long Q interested	111	13.33	86.19
Replacement, long Q non-interested	55	6.6	92.8
Replacement, short Q interested	11	1.32	94.12
Replacement, short Q non-interested	24	2.88	97
.	25	3	100
Total	833	100	

Type of sampling, surveyed short Q respondents

	Freq.	Percent	Cum.
Long Q, interested	11	1.17	1.17
Long Q, non-interested	2	0.21	1.38

Short Q, interested	189	20.06	21.44
Short Q, non-interested	472	50.11	71.55
Replacement, long Q interested	9	0.96	72.51
Replacement, long Q non-interested	6	0.64	73.14
Replacement, short Q interested	20	2.12	75.27
Replacement, short Q non-interested	195	20.7	95.97
.	38	4.03	100
Total	942	100	

In the sampling of short and long questionnaire respondents a total of 27 and 28 percent respectively had to be replaced with households identified as replacement households. There could be a number of reasons for replacement. These include the census lists provided by ADD not being completely updated, the household not being located based on the name of the household head alone, the household be unavailable at the time of the survey or the household or the household simply refusing participation.

Since we have no information on the non-response households we are unable to assess the bias in terms of observable characteristics of these households. We do, however, have qualitative information on why the sampled households had to be replaced, since the field supervisors had to note the reason for replacing a household in the first round of data collection.

While not being able to identify households who had indicated interest in participation (due to them recently signing up) this was indeed a problem for the households sampled based on the ADD census. Only few households refused participation, but some households were unavailable for the time of the survey. [A more thorough analysis on the exact reason for non-response is forthcoming.]

The response rate is taken into account in the sampling weights as described above.

4.4 Tracking

To minimize the bias arising from attrition special focus was given to tracking any households – or household members –that had moved between the different survey rounds.Tracking was only completed in the round 3, 2011 data collection.

It is important to note that tracking were done at the individual rather than the household level: I.e. if the individual who responded to either the household head or female questionnaire in round 1 had moved out, the person was tracked according to the rules set out below. If the household had split up, e.g. due to divorce, we have information on both the two new households. This was only done for the long questionnaire respondents. In the case where a short questionnaire respondent had moved out, the original household was surveyed in following rounds.

4.4.1 Rules for tracking

The following instruction was given to the supervisors in round 3 as to when a household had to be tracked (see also the front matter and tracking sections of the questionnaire for further information on the practical implementation of the tracking):

"When no designated respondent is available at the first household visit, enumerators will inquire about the availability of the designated respondent and try to schedule a date and time for the interview. If the designated respondent is still living within the household and is expected to return before the end of field work, enumerators will attempt to interview the designated respondent at a revisit to the household. A total of three visits will be made to interview the designated respondent. If the designated respondent is expected to return to the household within a couple of days, the survey team may revisit the household. If the designated respondent is not expected to return for a considerable period of time, but still before end of field work, the interview may be given to the tracking team.

When the designated respondent is expected to be away for the entire duration of the stay of the survey team in the village, or has left the household permanently, the enumerator will obtain information about the current location of the designated respondent on the tracking section of the long and short questionnaires." (Rockwool IKI ToR, June 2011)

If the designated respondent was not available for interview, information about the current location (a map and description of the current location, phone number of the designated respondent as well as the names of potential significant individuals in the new area) of the respondent was recorded. Furthermore, we collected information on whether the respondent was expected to return, and if so at which date, or whether the designated respondent had moved out of the household permanently.

Based on the information about the current whereabouts of the designated respondents, supervisors in collaboration with data collection managers from RFF decided whether the individual should be tracked or not. Tracking was in general done when the respondent was still living within approximately one hour's drive from the border of the TA Mwirang'ombe (the original survey area). This primarily meant tracking individuals who moved into neighbouring TAs or to the nearest larger cities, most notably Karonga[1]. Due to an unstable political situation at the time of the data collection affecting the larger city of Mzuzu it was decided not to track at this location (a location that is furthermore a two hour drive away) despite a number of designated respondents having moved here.

If the designated respondent could be located, two types of interview occurred; 1) If the designated respondent was still part of the original household, but was temporarily away (at the new location) the designated respondent was administered his/her questionnaire. 2) If the household had split up, the designated respondent was administered his/her questionnaire as well as the abbreviated version of the other long questionnaire. In this instance the other designated respondent (at the original household) was likewise administered his/her own questionnaire and the abbreviated version of the other questionnaire.

In a number of cases the designated respondent could not be located despite the new location being within the one hour driving distance. This could be due to inadequate information about the new location, the new dwelling being located but the designated respondent being away or due to the designated respondent changing name when re-marrying, in which case the local community at the new location would not know the designated respondent. When the designated respondent could not be locatedtwo

[1] NOTE: A map of the actual location of tracked households would be nice.

different scenarios were possible; 1) The designated respondent was still part of the original household. In this case, the designated respondent for the other part of the long questionnaire responded to an abbreviated version of the questionnaire in question. 2) If the household had split up, the 'new' household was not interviewed, but the designated respondent at the original household answered his/her own questionnaire as well as an abbreviated version of the other long questionnaire.

4.4.2 Extent of tracking

A total of 39 households were tracked in the 2011 survey round.

4.5 Attrition

Apart from the tracking described above, households were the designated respondents were not found at home were revisited a total of three times. If, at the third visit, the designated respondent could still not be interviewed, another household member was asked the parts of the questionnaire that contained household level information.

Despite this care in maintaining a balanced panel through careful revisits to the households surveyed in the first round in 2009, some attrition still occurred. The attrition in 2010 and 2011 will be discussed separately below along with the potential reasons for the attrition.

4.5.1 Round 2 – 2010

Households were not tracked outside the original survey area in the 2010 data collection. Still, the attrition for the long questionnaire respondents was reasonably low. A total of 799 households were interviewed in 2010. Compared to the original 833 interviewed households this amounts to an attrition of just 4,1percent.

Due to unfortunate events the data from the short questionnaires completed on PDAs was lost. This means we only have data from the short questionnaire interviews completed on paper: 218 observations in total.

4.5.2 Round 3 - 2011

In the final round of data collection more emphasis was put on tracking households and designated respondents who had moved. Thus the attrition rate was kept at a very low level, and some households that had not been interviewed in round 2 were located and interviewed for the endline. As explained in the tracking section above (and further elaborated on in the following section on split households) we gathered information on both new households in cases where a round 1 long questionnaire household had split up.

The datasets contains information on all 834 long questionnaire respondents and all 942 short questionnaire respondents. For the households that were not successfully tracked we only have information in the front matter and tracking sections. These households are still present in the remaining sections of the data but have missing values for all observations.

The round 3 long questionnaire data contains 834 observations with non-missing answers. Of the original 833 households we identified and surveyed 817 households (17 households were split between round 1 and round 3). The total attrition for the long questionnaires is 16 households – only 2.0 percent of the original sample.

The round 3 short questionnaire data contains 906 observations with non-missing answers. 8 of these are split households. Thus we have 898 of the original 942 households successfully interviewed, resulting in an attrition of 4.7percent.

4.5.3 Determinants of attrition

In the following, only the attrition from round 1 to round 3 will be considered. It is intended as merely descriptive of the baseline differences in observable characteristics of the households that could not be surveyed in round 3. Note that any attempt at tracking is not taken into consideration.

This has not yet been analyzed.

4.5.4 Split households

As described in the tracking section above we tracked any new households that had occurred due to a split of the original household. We defined a split as the household head or spouse of the household head moving out of the original household. Thus we did not track e.g. a daughter having moved out of the household to form a new household in another location.

In total we have 25 split households (17 in the long questionnaire and 8 in the short questionnaire data) where information on both the original household and the new household was recorded.

Table 5: Stated reason for split[2]

Reason for split	Freq.	Percent	Cum.
Illness	2	8.7	8.7
Marriage	4	17.39	26.09
Divorce	17	73.91	100
Total	23	100	

4.6 Problems encountered in collection of data

The following provides some comments regarding the collection and entry of data from the Invest in Knowledge Initiative (IKI). IKI delivered a field report following each round of data collection with information from the field staff on the ground that might be of importance for understanding the delivered data.

4.6.1 Round 1

Towards the end of data collection some interviewers had to leave the data collection. This created a need for including new personal late in the survey. These interviewers were quickly trained in conducting the short questionnaire, which was the only type of interview conducted by these late arrivals. There might, however, be a need for taking interviewer effects or the time of the interview into account when investigating the data.

It is also noted that the expectations section was the most difficult to complete – both for the interviewers and the respondents. The respondents "*picked up this [expecta-*

[2]Missing information for two of the split households.

tions] section after a few days in the field." It is noted that the expectations section caused specific problems in Nkhubasanga and Mwachesese villages.

There was also some problems identifying the designated respondents in the villages: "*Some names on the sample list could not be identified by villagers or they could be reported to be in a different village*". A number of households were also listed twice on the rosters.

Similarly, a number of designated respondents had to be replaced, since the respondents refused participation in the survey. IKI notes that "*Some designated respondents reported to have been promised to be given loans and noting that the field team had come for a different purpose, they were not very willing to respond to the questionnaires.*"

Due to these problems with identifying and interviewing designated respondents, not all villages ended up having 14 interested and 4 non-interest households interviewed using the long questionnaires, nor 7/14 interested/non-interested households interviewed using the short questionnaire.

It is also noted that local authorities (village headmen etc.) in general were cooperative, with the village of Chipemphe being a noteworthy exception, where the village officials were drunk through most of the survey period.

4.6.2 Round 2

A number of villages were noted as especially difficult to reach the respondents. These include Kapiyira, Maxwell, Kanyuka, Mphangwenjire, Kalimunda, Mwazorokele 1, Mwazorokele 2 and Mwera, which all had people working at the rice scheme in Vowve.

Although no actual tracking was done in round 2, the households that had moved within the survey area was attempted interviewed.

Some problems with the conducting the short questionnaire interviews were also noted, including malfunction of the PDAs and inconsistencies in the hard copy of the short questionnaire relative to the identical sections from the long questionnaires.

Finally, IKI provided some recommendations for future data collection: timely training in the PDAs, hiring of bicycles, timely preparation of questionnaires and providing incentives for the interviewees such as bars of soap or sugar.

4.6.3 Round 3

Some duplicate households were identified during the third data collection round. These households had been administered both a short and a long questionnaire in the previous rounds. In the third round, only the long questionnaire was completed.

IKI also notes some problems in tracking down respondents that had moved. Especially respondents that had moved away to (re-)marry were difficult to identify, as "*It was again difficult to trace some respondents especially females who have a tendency to change name whenever they re-marry somewhere else.*" Furthermore, "*There were also cases where the informants would not provide exact details that*

would lead the team to these respondents in such cases it was difficult to trace such respondents."

5 Description of questionnaire modules and variables

5.1 Questionnaires

The survey consisted of four different questionnaires: a village questionnaire, a short questionnaire and a household head and female questionnaire. The latter two were administered to the household head and his spouse from the same household. A household selected for surveying could thus either be administered a short questionnaire or a set of household head and female questionnaires as described above in the section "Sampling strategy". All questionnaires were developed in English and subsequently translated into the local language – Tumbuka – by IKI.

In the following sub-sections, the four different questionnaires are described in detail. It should be noted that the four questionnaires were not identical across rounds. Some sections were not included in some of the rounds. In the following description, each section reading is followed by information on the rounds in which the section was included. As an example "Household roster (R1, R2, R3)" indicates that the household roster was included in all three household rounds. "Front matter and tracking (R1, R2, R3*)" indicates that the section was included in all rounds, although the section was altered in the third round.

5.1.1 Household head questionnaire

The household head questionnaire was administered to the head of the household. The household members themselves identified the head of household. If the household head was the same as the designated respondent for the female questionnaire, or the household head was female, the two questionnaires were administered to the same individual, although care was taken not to conduct the two interviews back to back.

The household head questionnaire contained the following sections:

Front matter and Tracking (R1, R2, R3*):

These first sections gathered information about the respondent, the time and length of interviews, the GPS location of the household as well as information about the whereabouts of the respondent in case the respondent could not be located at the site of the household.

Household roster (R1, R2, R3):

This section gathered basic information about the members of the household, such as their age, gender, occupational status, health status etc.

Household members were identified using the following rules:

A HOUSEHOLD is one or more persons who have usually slept in the same dwelling and taken their meals together during at least three (3) of the twelve (12) months preceding the interview.

A DWELLING is the house, houses or apartment in which the household members are presently living.

However, the following people are household members, even if they have spent fewer than 3 months in the dwelling in the past 12 months:

- The person identified as the head of the household;
- Persons who just joined the household and expect to be long-term residents (i.e. expected to be residing in the household in the next 6 months), such as newborn infants aged less than three months or new spouses.

Furthermore, the following people are not household members, even if they have slept in the same dwelling and taken their meals with the rest of the household for the entire 12 months before the survey:

- Tenants and boarders, i.e. individuals who pay the household in cash or in kind (or in labor) for their own food and lodging. You will often need to probe to see if a person is a household member or a tenant, asking if they pay for their board (even if the payment is in kind or in labor).

General household characteristics

This section gathered information about the dwelling of the household as well as information about the household head, such as the religion, ethnicity and years living within the village.

Assets

Gathered information about the number of assets owned by the household. Both unproductive (e.g. TVs, tables, chairs) as well as productive (e.g. hoes, machetes) assets.

Spouses, parents residing elsewhere

This section was included to gather information about potential polygamous households. It gathered information on the names and place of residence of the (other) wife(s) of the household head, as well as any parents-in-law of the household head not residing within the household being surveyed.

Agricultural crop production

Gathered information about the most recent agricultural cycle, both the inputs into production (hired labor, seeds, fertilizer, manure), the harvest and whether the harvest was sold or used for own consumption.

Expectations

[Helene to update]

Storage

The purpose of this section was to understand how much of the preferred stable crop the household was storing, and how much they expect to be storing in January.

Risk aversion

This section asked the respondent to choose between two hypothetical crops. One would generate a certain yield, while the other involved a 50/50 chance of providing a small or large yield. The expected value of the latter would differ across the questions asked, in other to illicit the risk aversion of the respondent. Crops rather than monetary returns were used in the elicitation of risk aversion since agricultural crop production is the main source of income and thus also the main source of risk faced by the households in the area.

The questions were explained in the following was:

"Suppose that you receive a free gift of seeds for a small part of your land. You have the choice between two types of seeds of maize. Both need the same labour and fertilizer and taste the same. One type of maize yields 20 tins of maize for sure. The yield of the other one can change from year to year."

An example of the question asked was:

"Would you choose the crop that yields 20 tins of maize for sure, or the crop that yields a 50/50 percent chance to receive 40 tins but with a 50/50 chance has a yield of 0 tins?"

As the section was not made to force consistency in the answers, we included a 'trick' question in the 2011 data collection:

"Would you choose the crop that yields 20 tins of maize for sure, or the crop that yields 20 tins for sure and a 50/50 change to receive another 20 tins?"

If the respondent for this question chose the crop with a certain 20 tin yield, the interviewer was instructed to explain the concept to the respondent once again.

Intertemporal discount rate

The aim in this section was to learn whether the designated respondent preferred having something now or prefers to wait for the same thing (or something different) in the future.

A sample question is:

"Would you prefer to have the 2000 MK now or 2800 in one month?"

Unlike the risk aversion section, the skip patterns in this section forced consistency in the answers of the respondent.

Interseasonal compensation

The purpose of this section was to get some understanding about how much food the household would be willing to give up now in return for more or less food in January, the typical hungry season in the area, just before the green harvest.

Wage income

This section asked about the wage income of any household members during the past 12 months. Wage income could be both in cash and in kind and any length of employment was included. Information was gathered also on the location and type of occupation.

Risk response

This section assessed the shocks which the household had experienced within the past five years. Furthermore, it asked the household to assess which types of risks they considered to be most threatening to their future wellbeing.

Psycho-social variables and general outlook (R2, R3):

This section assessed the general self-efficacy of the respondent[3] as well as the general happiness of the respondent and his/her view on the future of the household.

Savings groups (R2, R3):

This section specifically asked about the household's involvement in the VSLA in the area.

Spending money (R3):

The aim of the section was to gain an understanding of the amount of spending on 'irrelevant purposes', such as snacks, alcohol, gambling and prostitution. Since these were very delicate questions, care was taken that the respondent was alone at the time of asking the questions.

Finally, the section concluded by some questions assessing the cognitive abilities of the respondent. A series of mathematical questions of increasing complexity were asked.

5.1.2 Female questionnaire

The female questionnaire involved some sections similar to those in the head questionnaire. In case a single individual answered both questionnaires, an abbreviated version of one of the questionnaires was used, to avoid asking the same questions twice. The guidance to the interviewer for which sections to administer in this case was described on the front matter of both the household head and female questionnaire. The following sections were included in the female questionnaire. Where the section is analogous to the sections described in the head questionnaire, no further description is given here.

Front matter and tracking (R1, R2, R3

Livestock

[3]Taken from Schwarzer and Jerusalem (1995)

We asked about the current livestock holdings of the household, as well as the purchase and sale of livestock within the past 12 months.

Income sources

This section assessed the types of income sources of the household, as well as identifying the household member responsible for the specific income sources.

Household enterprises

The profits, inputs and working capital of any household enterprises, from fishing over petty trade to actual production of goods.

Health

This section assessed whether any household members had been seriously ill during the past year and the actions taken by the household.

Consumption expenditures

Household food consumption during the past week prior to the interview as well as consumption of a few other items such as soap, clothes and alcohol over a longer time period. See the section on "Creating aggregate consumption" below for a detailed description of this section.

Household food security

This section contained questions about the experienced food security during the past 12 months – i.e. the number and size of meals as well as the composition of meals in different months within the past year. The section also asked hypothetical questions about the expected food security in scenarios where either the household alone or the entire village were experiencing a poor harvest.

Psycho-social variables and general outlook

The female questionnaire included a few extra personality traits compared to the head questionnaire.

In the 2009 questionnaire, questions on self-efficacy and locus of control[4] were included. In the 2010 questionnaire, the locus of control questions was not included, but a series of questions assessing the self-esteem were included instead. In the 2011 questionnaire both self-efficacy, locus of control and self-esteem questions were included.

In all the rounds, questions about the general happiness and accomplishment of the respondent as well as the outlook on the future of the household were included.

Intertemporal discount rates

Expectations

[4]Using a subset of the Rotter (1966) questions.

Membership in groups

This section was included in all rounds, and asked for any groups the household may belong to. This could be microfinance groups, agricultural groups, village boards etc.

Savings groups

Network roster and transfers

This section asked the household to list all the individuals outside the household the household could rely on, or which relied on the household in times of hardship. Subsequently a series of questions were asked about the depth of this network – i.e. the actual transfers between household in the past year as well as the potential transfers.

Spending money

Credit

The aim of this section was to know both the actual and potential credit sources of the household.

Savings

This section assess the amount of savings, either monetary or in gold etc. by the household.

Risk aversion

Interaction with survey respondents

For each village, a list of the respondent household heads was compiled. Each surveyed household was then asked about their relations to each of the households on this list.

Child anthropometry

A standard set of questions on the age, height, weight and MUAC of each household member under the age of seven.

5.1.3 Short questionnaire

This questionnaire was an abbreviated version of the household head and female questionnaires. The short questionnaire only included questions on the central variables expected to be used for the impact assessment of the VSLA project.

The following sections were included in the short questionnaire:

- Front matter and tracking
- Household roster*
- General household characteristics*
- Income sources
- Consumption expenditures*
- Household food security*

- Agricultural production*
- General outlook
- Assets
- Membership in groups*
- Savings*
- Interaction with survey respondents

The asterisk marks the sections that were included in an abbreviated form compared to the sections from the household head and female questionnaires.

5.1.4 Village questionnaire

The village questionnaire was designed to assess the physical attributes of the village. In order to get as accurate a picture as possible, the village questionnaire was completed by interviewing a group of village leaders simultaneously. At least four knowledgeable respondents (religious leaders, health workers, headmasters etc.) should be present, including the village headman, at the time of the interview.

The village questionnaire contained the following sections:

Front matter

Information about the interview conducted as well as GPS-locations of important features within the village; the village center, the nearest market, the nearest tarmac road, nearest primary and secondary school and nearest health clinic.

Respondent characteristics

Basic information about the respondents present at the time of the interview.

Village characteristics

This included basic information about the village, such as the size (number of households, approximate number of acres), the number of female headed households, the number of orphans etc.

Village infrastructure and organisations

This section gathered information on whether the village has any schools, bus stops, microfinance institutions, banks etc. If the village did not have a specific unit of infrastructure, the distance to the nearest one was recorded.

Economic activity

The aim of this section was to gather information about migration work, agricultural activity, information about the distribution of coupons for seed and fertilizer, as well as price information on agricultural piece-work.

Village history

This section asked for information about any shocks that had occurred in the village within the past five year (in the 2009 questionnaire) or the past year (in the 2010 and 2011 questionnaires).

Price information

This section contains information about the seasonal variability in the prices of the most important types of food, as well as the price of credit from moneylenders.

Individual interviews

This section asked a number of questions to each of the respondents individually. The section contained questions on their individual assessment of the prices in the village, as well as abbreviated versions of the psycho-socials and spending money sections from the head questionnaire (provided the respondents for the village questionnaire had not already been subject to the head questionnaire).

5.2 Questionnaire modules

See section 8.6.3 for a table presenting sections, number of observations and unit of observation.

6 Additional datasets created

Beyond the individual sections of the questionnaire (and the panel version of these datasets) a couple of datasets have been created for the ease of the end user. The contents of these will be described below.

6.1 All purpose panel cleaned dataset (APPCD)

[Update with the variables contained as well as their exact definition]

7 Created variables

Based on the data collected a number of more aggregate variables were created. These are of a more general nature, and might be of use in any subsequent analysis: aggregate consumption measures, an estimated adult equivalent number of household members and two different poverty indicators: PAT and PPI. The current section describes the variables and their exact definition. All these variables are available in the APPCD dataset described above.

7.1 Adult equivalent

The size of a household is important in assessing the income per individual. A simple count of the number of member in a household is one possibility, but this does not reflect the difference in e.g. calorie intake between an infant and an adult.

To take the gender- and age-composition of the households into account when generating household level income and expenditure measures we calculate the number of adult equivalent living in each household following Skoufias et al. (1999).

Table 6: Adult equivalent conversions

Age (years)	Male	Female
Under 5	0.41	0.41
5-10	0.80	0.80
11-14	1.15	1.05
15-19	1.38	1.05

20-34	1.26	0.92
35-54	1.15	0.85
55+	1.03	0.78

The adult equivalent size of the household is especially important in generating the consumption per household member as will be described below.

In round three, gender info for household members was pre-coded on questionnaires using data that was not re-entered. Moreover, the exact dataset used for pre-coding was not saved. In constructing adult equivalent, the gender info is needed and thus, we had to assume that the individual IDs used in R1 and R3 are the same. This assumption is thoroughly tested and found valid in the do-file Cleaning H3A.do.

7.1.1 Aggregate consumption

In a survey such as the LSMS surveys conducted by the World Bank, one of the primary objectives of the data collection is to produce a very accurate measure of the income and expenditure level of the households in order to assess the extent of poverty. This results in a very elaborate consumption expenditure section. In the Malawi Second Integrated Household Survey (IHS-2) carried out by the World Bank in 2004/05 for instance the expenditures for a total of XX items was included. This was not feasible in our data collection.

As a proxy for the total food consumption we identified the most important items for the households in the Karonga district and included only these items in the questionnaire.

In an attempt to construct an aggregate consumption measure including the remaining food items and the non-food consumption component a scaling has been implemented. First the consumption measure is scaled by 1.123596, corresponding to a scaling by 11 percent. Next, the aggregate food consumption measure is scaled by 1.58 which is the ratio by which food is consumed to non-food. This information is gathered from the LSMS dataset "ihs2_pov.dta" and is region specific. The obvious problem with this sort of scaling is the assumption of identical consumption pattern across all households irrespective of consumption level. The name of this variable is "aecons_4p".

7.1.2 Items included in questionnaire

The 17 (18)[5] items included in the questionnaire were selected based on the 2004/05 Malawi Second Integrated Household Survey (IHS-2) carried out by the World Bank. Using this dataset, we identified the primary items of consumption for the Karonga District and Mwirang'ombe TA. Since there were only 35 observations within Mwirang'ombe TA, it was decided to use the top 17 food items in Karonga district, measured by the value of consumption.

Table 7: Top foods in Karongadisctrict, measured as share of total food expenditure

[5] In round 2 and 3 an "Other beans" category was added to the questionnaire. Due to this, two different measures are created in R2 and R3 – only including the other excluding this additional category.

Rank	Item #	Description	Share	Cummulative
1	503	Fresh fish	0.2614	0.2614
2	502	Dried fish	0.1726	0.4340
3	102	Maize ufa refined (fine flour)	0.0956	0.5296
4	202	Cassava flour	0.0762	0.6058
5	101	Maize ufamgaiwa (normal flour)	0.0481	0.6539
6	404	Nkwani	0.0420	0.6959
7	105	Green maize	0.0273	0.7232
8	106	Rice	0.0254	0.7486
9	810	Salt	0.0221	0.7707
10	302	Bean, brown	0.0219	0.7926
11	803	Cooking oil	0.0181	0.8107
12	507	Chicken	0.0178	0.8285
13	201	Cassava tubers	0.0165	0.8450
14	405	Chinese cabbage	0.0145	0.8595
15	801	Sugar	0.0128	0.8723
16	203	White sweet potato	0.0096	0.8819
17	408	Tomato	0.0093	0.8913

As can be seen from table 1, these 17 items make up 89 percent of total food consumption in the Karonga district from the 2004/05 Malawi HIS-2. For the subsample of Mwirang'ombe, the 17 items make up 91 percent of total food consumption, although an item as "wild green leaves" which made up 1.8 percent of total food consumption for the Mwirang'ombe TA subsample was not included.

7.1.3 Units of measurement

Again following the IHS-2 we allowed for 19 different units of measurement when the households had to state their food consumption in the past week. The 19 measurement units were as follows:

Table 8: Measurement units used in the 2009 data collection

UNIT CODE
1. Kilogramme
2. Gram
3. Pail (small-5littres)
4. Pail (large-25litres)
5. No. 10 plate
6. No. 12 plate
7. Bunch
8. Piece/Cob
9. Heap
10. Bale
11. Basket (*Dengu*) shelled
12. Basket (*Dengu*) un-shelled
13. Liter

14. Milliliter
15. Basin/pot
16. Cup
17. Spoon
18. Tin
19. Other (specify)

By using these we could rely on the conversions into grams for each unit-item combination collected by the World Bank in the IHS-2. One notable exception is the 'basin/pot' measurement unit. This was not part of the IHS-2 questionnaire. How this is dealt with in each of the methods used is described below.

However, due to problems in these conversions (as will be described below), we used a different approach in the 2010 and 2011 data collections. Here we only allowed for four different unit codes (as seen below), but combined these with the picture unitsseen in appendix A.

Table 9: Measurement units used in the 2010 and 2011 data collections

UNIT CODE
1. Kilogramme
2. Gram
3. Litre
4. Millilitre

7.1.4 Calculating value of consumption in 2009 dataset

We calculate the total value of food consumption by using the specific item-unit prices collected in our survey. While we used the measurement units described in table X above based on the IHS-2, the resulting measure of consumption in our sample is implausible. This is primarily due to problems with a couple of the unit conversions. A 'tin', for instance, contains 33 grams of maize according the IHS-2 conversions. However, a tin of maize was discovered to contain between 6 and 13 kilos, depending on the type of tin.

Instead we use item-unit specific prices. In practice, we have households who have stated the price paid for the quantity of e.g. maize bought in tins, some who stated the quantity bought quantity in kilogram, and some who stated the quantity bought in pails. This allows us to generate the median item-unit specific prices and evaluate the household consumption by these prices. Due to the relatively large number of item-unit prices to calculate, these are done across the entire survey area. The cost of not being able to take variation in prices within the survey area into account was considered acceptable compared to the problems with the IHS-2 conversions.

The item-unit combinations for which we have no information on the price we use the median value of consumption per adult equivalent for the specific item, weighted by the number of adult equivalent members of the household. There are a total of 283 observations for which we lack information on the price of one of the consumption items. Furthermore, some of the item-unit prices are based on a relatively low number of observations.

The following tables provide some descriptive statistics (un-weighted)on the final measure of food consumption per adult equivalent[6].

Table 10: Mean value of consumption (2009 MK) for each item and by type of questionnaire[7]

Item	All Mean	Long Mean	Short Mean
Normal flour	38	38	38
Fine flour	87	93	82
Green maize	6	5	7
Rice	57	61	54
Cassava tubers	22	21	22
Cassava flour	34	36	32
Sweet potatoes	17	16	19
Brown beans	19	19	19
Any other beans	.	.	.
Pumpkin leaves	5	5	5
Chinese cabbage	5	5	5
Tomato	30	30	30
Fresh fish	41	47	36
Dried fish	26	25	27
Chickens	63	59	67
Sugar	45	32	57
Cooking oil	28	28	28
Salt	15	11	18
Total	539	530	546
#Obs	1709	822	887

Table 11: Consumption per adult equivalent round 1 (2009 MK)

Item	# Obs.	Mean	Min	Max	Median
Normal flour	1,709	38	0	583	0
Fine flour	1,709	87	0	1,256	81
Green maize	1,709	6	0	388	0
Rice	1,709	57	0	887	35
Cassava tubers	1,709	22	0	1,258	9
Cassava flour	1,709	34	0	552	17
Sweet potatoes	1,709	17	0	1,188	0
Brown beans	1,709	19	0	735	0
Any other beans	1,709	.	.	.	.
Pumpkin leaves	1,709	5	0	256	0
Chinese cabbage	1,709	5	0	272	0
Tomato	1,709	30	0	1,500	23

[6]See the section below on the creation of the adult equivalent measure.
[7] Taken from "Valconsip by item and type of questionnaire.xlsx"

Fresh fish	1,709	41	0	701	26
Dried fish	1,709	26	0	333	18
Chickens	1,709	63	0	1,351	0
Sugar	1,709	45	0	23,077	28
Cooking oil	1,709	28	0	462	20
Salt	1,709	15	0	2,521	8
Total HH consump-tion	1,709	539	0	23,322	443

7.1.5 Calculating calorie intake 2009

This is currently not done as the calculation of consumption expenditures in 2009 does not involve calculating the consumed amount in kilogram of each food item.

7.1.6 Calculating value of consumption in the 2010 and 2011 datasets

For the 2010 and 2011 consumption modules, the units of measurement were changed (as described earlier). This made the conversion into kilogram easier as we collected our own conversion measures. To be as consistent across rounds as possible (despite the different measurement units) we calculate the prices for the 2010 and 2011 datasets in the following way:

- Estimating area-specific prices per kilogram by:
    1. Converting all purchased quantities into kilograms using our collected unit-conversions
    2. Estimating the median price per kilogram in the entire survey area
    3. Estimate the value of consumption for all households by evaluating the amount consumed at the calculated median prices

Similar to the consumption measure calculated for the 2009 dataset this does not allow us to capture any geographical variation in the price of food items.

Despite the interviewers being instructed in collecting the amounts consumed in the picture units for each specific item of consumption, a few cases still exist where a household reported consumption of a food item in a picture unit for which a conversion was not collected. For these we simply replace with the median expenditure per adult equivalent in the entire survey area weighted by the number of adult equivalent members of the household.

The following tables provide some descriptive statistics on the consumption measure generated for round 3 data.

Table 12: Mean value of consumption (2009 MK) for each item and by type of questionnaire

	All	Long Q	Short Q
Normal flour	12	11	14
Fine flour	118	105	130
Green maize	6	5	6
Rice	97	88	105

Cassava tubers	16	16	16
Cassava flour	44	41	46
Sweet potatoes	37	34	40
Brown beans	12	11	13
Any other beans	14	14	13
Pumpkin leaves	17	12	21
Chinese cabbage	6	5	7
Tomato	42	34	50
Fresh fish	51	45	57
Dried fish	70	67	73
Chickens	64	59	70
Sugar	47	45	48
Cooking oil	16	11	20
Salt	10	10	11
Total	684	620	743
# Observations	1738	832	906

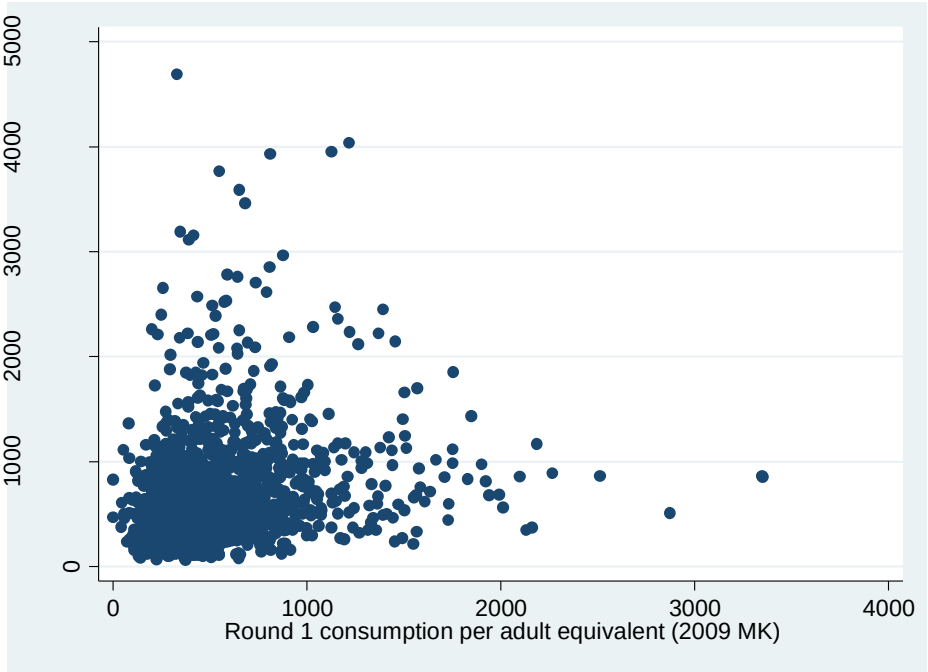
Table 13: Consumption per adult equivalent round 3 (2011 MK)

	Mean	Median	Min	Max
Normal flour	13	0	0	348
Refined/fine flour	118	106	0	997
Green maize	6	0	0	413
Rice	98	65	0	1,240
Cassava tubers	16	0	0	477
Cassavaflour	44	11	0	1,512
White sweetpotatoes	37	17	0	1,284
Brown beans	12	0	0	301
Any otherbeans	14	0	0	463
Pumpkinleaves	17	8	0	1,505
Chinesecabbage	6	0	0	258
Tomato	43	28	0	1,553
Freshfish	51	27	0	1,002
Driedfish	70	21	0	18,635
Chickens	65	0	0	1,111
Sugar	47	41	0	848
Cooking oil	16	0	0	802
Salt	10	8	0	93
Total	684	547	58	20,661
# Observations	1,738	1,738	1,738	1,738

7.1.7 Correlations between the 2009, 2010 and 2011 value of consumption

The following descriptive statistics indicate the correlations between the round 1 and round 3 consumption measures. Note, however, that the intervention was intended to affect the consumption levels of the households.

Figure X: Consumption per adult equivalent in round 1 and 3



7.1.8 Converting consumption expenditures to USD

To calculate the USD-equivalent of food consumption we use the poverty-weighted purchasing power parity (4P) measures calculated by Deaton and Dupriez (2011). The advantage of the extra 'P' lies in the basket of goods used to calculate the prices across countries. Where PPP uses an economy wide basket of goods, the PPPP approach uses a basket of goods typical for the poorest consumer. As such, the measure should more adequately reflect the price changes experienced by the poorest parts of the population.

In practice we use the following conversions to convert all prices into 2005 USD.

Year	Malawi Kwacha per USD (4P)
2009	91.19
2010	97.94
2011	105.41

7.1.9 Suggestions for improvements

The 2009 measure is somewhat problematic, as it is based on subjective assessments on the validity of the conversions. Thus, a suggestion for an improved measure could be to collect all the conversions needed to calculate an entirely new measure.

7.2 Converting R1 agricultural production into kilograms

In R1 agricultural production is measured in 14 different unit codes covering kilograms, tins, baskets etc. Due to awareness that the unit code tin covers a variety of sizes it was decided to use picture codes for R2 and R3. In order to make data comparable across all three rounds following rules were used for converting R1 agricultural production into kilograms:

Unit code 1-3: All measures are in kilograms (respectively 1 kg, 50 kg and 90 kg) and conversion is therefore easily done. This yields 994 observations corresponding to 51 pct. of the harvest answers.

Unit code 4-5: Both measures are in liters and here LSMS conversion data is used to convert into kilograms. LSMS assume 1 liter to equal 1 kilogram for all crop codes. This yields 45 observations corresponding to 2 pct. of the harvest answers.

Unit code 13: Picture codes from R2 were used to convert tin into kilograms. The assumption is that a given household is pointing at same size of tin in R2 as thought of in R1. Data is obtained for 224 observations, while other 262 observations are not matched. These are imputed using the rule explained below.

For the remaining missing observations LSMS conversion data is used. This gives 554 observations of which 51 pct. come from cotton and burley tobacco (both usually measured in bale).

Now 121 harvest answers are missing in R1 stemming primarily from tomatoes and cassava and 262 tin observations are missing due to no match. These observations are all imputed by the following rule:

1. Calculate median harvest per acre for all crops separately
2. Impute the missing values by the median times acres used for the specific crop

This leaves only four missing observations of tot_harvest (measured in kg) given a positive answer of agricultural production.

7.3 PAT and PPI

Poverty Assessment Tool (PAT) and Progress out of Poverty (PPI) are two proxy indicators that use a limited number of questions from the LSMS surveys to predict either consumption (PAT) or probability of being poor (PPI). PAT and PPI were constructed following legislation by the US Congress to better monitor targeting of US support microenterprise development (US Congress 2002). The goal was to construct indicators which are quantitative, objective and low cost.

Both PPI and PAT are computed from existing LSMS surveys and they both select a small number of variables and estimate parameters for these variables. These parameters can then be used to predict poverty using only the few selected questions.

The key difference between PPI and PAT is that PAT predicts total local currency consumption per capita, whereas PPI predicts the probability of being poor. PAT uses one of four methods to predict total consumption: OLS, Quantile, Linear Probability, and Probit. In our case, parameters were estimated using quantile regression at the 1.25USD poverty line.

Below we explain how exactly the two are constructed and particularly how we have dealt with following two primary challenges:

1) Variables not included in round one. Some variables were not included in the round 1 questionnaire, the most important being number of goats. This was simply a mistake in the questionnaire design. For these we used recall.

2) Missing observations. Since many of the variables are themselves combinations of other variables, the final measures initially had quite a large number of missing values. In some cases, we have imputed these with median values, while keeping track of the effect this had on the final measure.

7.3.1 PAT

As mentioned, natural logarithm of total consumption is estimated using quantile regression at the 1.25 USD poverty line. The data is the 2004 Malawi LSMS. For more information on the methodology, please see USAID 2012.

The parameters were provided by Anthony Leegwater at IRIS on Feb 23, 2010. They are listed in the "Coef." column in Annex X.

7.4.1.1 Recalled values

Recalled variables are listed in the table below. To check validity, we included recalls even though the questions were asked in the rounds that were recalled in two cases: goats and bicycles. Results are listed below.

Variable	Number of people in R1 with value > 0	Recall validity Share of respondents with the same recall as real
Flush toilet	8	
Coffee table	647	
Bed	1246	
Iron	552	
Goats	741	55% from R3 to R2.
Bicycle	Used for validity check only	69% (R2 recalled to R1) and 70% (R3 recalled to R2

7.3.1.2 Imputations

A maximum of five Imputations were made for each household, and we checked the effect of the imputations on the final measure by imputing with the 25th, 50th and 75th percentiles and comparing the three.

The number of imputations is listed below.

Variable	Number of imputations
pat2_hhsize	17
pat3_headage	48
pat4_hhsingle	6

pat5_litady_chi	10
pat6_mbedunone	18
pat7_nrooms	3
pat8_floorcement	1
pat9_toiletflush_RECALL_R1	82
pat10_electricity	7
pat11_bed_RECALL_R1	83
pat12_coffeetable_RECALL_R 1	83
pat13_iron_RECALL_R1	83
pat14_player	8
pat15_bicycle	2
pat16_powdersoap	50
pat17_dimba	3
pat20_nowngoats	47
pat21_headage2	48
pat22_hhsize2	17

The three percentile-imputations' averages and medians were:

Percentile	Control group	Treatment group	Total
25th	Mean: 4280 Median: 2761	Mean: 3415 Median: 2769	Mean: 3843 Median: 2766
50th	Mean: 4290 Median: 2772	Mean: 3422 Median: 2774	Mean: 3851 Median: 2772
75th	Mean: 4344 Median: 2835	Mean: 3467 Median: 2832	Mean: 3901 Median: 2835
Differences in means			
Differences in medians			

7.3.1.3 Correcting PAT-variables

In the PAT documentation form 2010 we originally relied on three variables were wrongly defined. "HH head is single" was really "Household head never married", "number of adult HH members (age>=18) who can read in Chichewa" was really "Share of adults who can read" and "Number of goats" was really "Household owns one or more Goats." The changes has been implemented without any problems, except for "household head never married" which we do not have data for in round 3. Since only two households answered "never married" in round 1 (2 of 830) and only one of these households answered "never married" in round 2 (1 of 831), we solved the missing data problem in round 3 by imputing everyone to zero.

7.3.2 PPI

The PPI measure is typically constructed using ten parameters, also from the LSMS. Contrary to PAT, the PPI parameters are from a model that predicts the probability of being poor using a probit.

In our case, we use only nine parameters. This is due to a mistake in the original PPI measure. They include an agricultural variable, which was not measured the right way. In an email from 03/06/2010, Mark Schreiner confirmed that there were indeed issues with the variable. He then recalibrated PPI with only nine parameters (see email from 15/6/2010), where he also attached the calibrated scorecard which we use in the data.

Later, on 11 January 2011, Mark rebuild the PPI. Since we were already back then working on the data, we choose to continue with the simple recalibrated data.

Emails documenting this are in the folder "Dropbox\VSLAproject\Documents\Other docs\PAT-PPI documentation\" as well as the subfolder \PPI Error\.

References

USAID 2012, Poverty Assessment Tool Accuracy Submission USAID/IRIS Tool for Malawi, downloaded December 2012 from http://www.povertytools.org/countries/Malawi/Malawi.html .

7.3.2.1 Imputations

A maximum of two imputations were made for each household. The number of imputations is listed below.

Variable	Number of imputations
ppi1_children_points	0
ppi2_water_points	14
ppi3_cooking_points	0
ppi4_light_points	10
ppi5_paraffin_points	0
ppi6_furniture_points	88
ppi7_bicycle_points	0
ppi8_player_points	4
ppi9_iron_points	87

8 Using the data

The following sections provide practical guidelines for installing and using the data, as well as some basi information on the steps undertaking by the research team in

arriving at the cleaned data. For a more detailed description of the cleaning process see the document XX.

8.1 How the raw data file is created

This section concerns how the data file is created and not what to do in order to use the data. For the latter see next section. First, the folder "Folder structure (empty)" is copied to the disc. This ensures that the folder setup is identical to the one needed in order to run the do-files. What is left to do is to copy the relevant raw data to the disc. The following lists the files to copy from the three rounds respectively.

R1:

- The files in the "raw" folder (22 files)
- The "cleanup" folder and its contents (3 files)
- The files in the "Sampling" folder (10 files)
- In the "Sampling" folder: The "DTA Log Sheet Completed Villages" folder and its contents (47 files)
- In the "Sampling" folder: The "CSV" folder and its contents (91 files)

R2:

- The files in the "export" folder
- The file "R2 DATA_VILLAGE QX_3005" which is directly in the pre_raw2 folder.

R3:

- The "shortqbatches" folder and its contents
- The "longqbatches" folder and its contents

Now the raw data file is complete. This is the only data file needed to run the do-files from the dropbox.

8.2 Guide to install and use data

Unzip the zipped folder "MalawiKAW" and save it somewhere on your computer. Apart from the folders containing raw data the other folders are empty.

Open the do-file "runall global.do" in the dropbox folder "…Dropbox\VSLAproject\Data". Now you need to define your own file paths as follows:

The global macro "MalawiGlobal" should refer to the unzipped "MalawiKAW"-folder. The global macro "dropbox" should refer to dropbox on your computer.

That is all. Now you can run the two globals just created and then section B, C and D in the same do-file. The runall-files clean the raw data and place the cleaned files in the folder called “cleaned” for the respective round of data.

8.3 Principles in cleaning the data

The causal structure from raw data to cleaned data and finally panel data can be described as follows:

- Raw data files, e.g. “A3A.dta” placed in “raw_r3”. These are raw data.
- Cleaning in progress files, e.g. “A3A.dta” placed in “cleaning in progress_r3”. These are copies of raw data to be used for cleaning.
- Cleaned data files, e.g. “A3A.dta” placed in “cleaned_r3”. These are the cleaned data files produced by the runall do-file.
- Panel files, e.g. “Panel AA.dta” placed in “constructed data_r3”. These are section panels using R1 and R3.

8.3.1 Which variables to impute/change?

Changes to the data have been done as little as possible and only where we have confidence that values are wrong other than good judgment. One exception to the rule is when the skip-pattern in the questionnaire is not followed resulting in answers on a variable that shouldn't have been answered. In this case, a value is imputed to missing.

8.4 Identification of households

In round 1, vid and hhid uniquely identifies the observations. However, in round 2 and especially round 3, vid and hhid no longer necessarily uniquely identifies the observations. For instance, the designated respondents constituting a round 1 household may have divorced, resulting in two separate households in round 3.

To create unique identification of the observations in each separate round the following new variables have been created:

- *r3hhid*: The household ID of the current household to which the actual respondent belongs. For households that have split up, the location of the original household has been assigned *r3hhid=1000+hhid*, while the new household has been assigned *r3hhid=2000+hhid*. For householdsthat have not split up, *r3hhid=hhid*.

8.5 Generating new variables

In some sections the cleaning file generates new variables. In A3AK anthropometric measure variables are created. In A3B variables on income sources are created. In A3B2 variables of totals and dummies are created. In A3FRISK risk aversion measures are generated. In A3HPART1 variables of value of consumption, calorie intake and consumption per adult equivalent are created. In A3L membership variables are generated. In a3lb membership variables are generated. In A3ME dummy for residence is generated. In H3FPART2 agriculture variables are created. In H3FRISK risk aversion measures are generated. In SA3B income variables are gen-

erated. In SA3H consumption variables are generated. In SA3I meal variables are generated. In SH3F agriculture variables are created. In SH3SAVE1 variables on savings are created.

8.6 Creation of panels

The panel datasets are currently only created using the round 1 and round 3 data. Round 2 data is still to be included in the panel datasets.

Panels for each section are created by do-files contained in the folder "Do-files_panel". Each do-file is named "Panel XX" where XX is name of the specific section, e.g. AB. Using section AB as example, the structure of the do-files can be described as follows: First the cleaned data file for section AB in R1 is loaded and variable names are changed in order to match R3. If same section is present in short questionnaire as well, variable names are changed here too. The modified datasets are saved as temporary files. Cleaned data for section A3B in R3 is loaded and the "3" is removed from all variable names. Now variable names are identical and constructed temporary files are appended onto R3 data. Finally, a dataset of attrition information is merged and the panel is saved as "Panel AB" in "$panel\Panel `section'.dta".

8.6.1 Questions that are not included in both rounds

Only variable names of variables present in both rounds are changed. Hence, variables which only are present in either R1 or R3 will be present under their original variable name in the panel dataset. Two exceptions are made: When variable names in R1 are identical to R3 but the questions are different, the variable names of R1 have been changed to either "variable name_R1" or the section name followed by a number not yet taken.

8.6.2 Sections that are not in both rounds

No panel is constructed when sections are not present in both rounds.

8.6.3 Overview of questionnaire sections and associated panels

Heading	Panel data	Questionnaire sections			Number of observations			Number of households			Unit of observation
		R1	R2*	R3	R1	R2*	R3	R1	R2*	R3	
FEMALE											
Health	Panel AA	aa	-	A3A	833	-	850	833	-	850	Individual
Child Anthropometry	Panel AAK	aak	a2ak	A3AK	1106	1249	1436	632	644	673	Child name
Income Sources	Panel AB	ab	a2b	A3B	13328	12000	12526	833	800	850	Income source
Livestock	Panel AB2	ab2	a2b2	A3B2	8330	7992	8356	833	801	850	Livestock type
Household enterprises	Panel AC	ac	a2c	A3C	1007	1235	2391	833	798	850	Business type
Transfer Networks	Panel AME	am	a2me	A3ME	6887	7474	9334	831	801	850	Network Roster ID
Spending money	Not created	-	a2f	A3F	-	741	850	-	740	850	Individual
Risk Aversion	Panel AFRISK	afRISK	-	A3FRISK	833	-	850	833	-	850	Individual
General Outlook Questions	Panel AG	g	a2g	A3G	833	743	850	833	743	850	Individual
Psycho-Social Variables	Panel AG2	ag2	-	A3G2	833	-	850	833	-	850	Individual
Consumption expenditures	Panel AH_part1	ah	a2h (not cleaned)	a3hpart1	16660	23151	17551	833	801	852	Consumption item
	Panel AH_part2	ah	a2h (not cleaned)	a3hpart2	16660	23151	3274	833	801	850	Consumption item
	Panel AH_part3	ah	a2h (not cleaned)	a3hpart3	16660	23151	3352	833	801	850	Consumption item
Household Food Security	Panel AI	ai	a2i	A3I	833	801	850	833	801	850	Individual
Intertemporal Discount Rates	Panel AID2	afRISK	A2ID2	A3ID2	833	743	850	743	801	850	Individual
Credit	Panel AJ_PART1	aj_part1	A2J	a3jpart1	833	774	850	833	745	850	Individual
	Panel AJ_part2	aj_part2	A2J	a3jpart2	107	774	239	107	745	239	Individual
Savings	Panel AK_part1	ak_part1	a2k_part1	A3KPART1	833	6913	7442	833	769	850	Saving type
	Panel AK_part2	ak_part1	a2k_part1	A3KPART2	833	6913	850	833	769	850	Individual
	Panel AKPART3	ak_part2	a2k_part2	A3KPART3	4165	991	1046	833	768	850	Borrower ID
Membership in groups	Panel AL	al_part1	a2l	A3L	9996	9552	10024	833	796	850	Membership group
Savings groups	Panel ALB	al_part2	a2lb	A3LB	833	801	851	833	801	851	Individual
Designated respondent tracking	Not created	-	h2track	A3TRACK	-	25	850	-	25	850	Individual

HEAD	Panel data	Questionnaire sections			Number of observations			Number of households			Unit of observation
		R1	R2*	R3	R1	R2*	R3	R1	R2*	R3	
Household Roster	Panel HA	HHroster	h2a (not cleaned)	H3A	5003	5995	5696	833	802	844	Household members
General Household characteristics	Panel HB	hb	h2b	H3B	833	799	844	833	799	844	Household
Assets	Panel HC	hc	h2c	H3C	833	799	844	833	799	844	Household
Spouses, parents residing elsewhere	Not created	hd_part1 and hd_part2	h2d_part1 and h2d_part2	H3D	Ambiguous	Ambiguous	844	Ambiguous	Ambiguous	844	Individual
Wage income	Panel HE	HE	h2e	H3E	909	948	965	833	802	828	Wage income
Agricultural crop production	Panel HF_part1	hf_part1A	h2f_part1	h3fpart1	833	2013	844	833	787	844	Individual
	Panel HF_part2	hf_part1B_new	h2f_part3_new	h3f_part2	2305	4649	1975	823	802	811	Crop ID
Risk Aversion	Panel HFRISK	hk	Not included	H3FRISK	833	Not included	844	833	Not included	844	Individual
Risk response	Panel HHpart1	hh	h2h	H3Hpart1	1237	1086	844	673	787	844	Individual
	Panel HHpart2	hh	h2h	H3Hpart2	1237	1086	880	673	787	581	Individual
General outlook	Panel HI	HI	h2i (not cleaned)	H3I	833	802	844	833	802	844	Individual
Intertemporal discount rates I	Panel HID2	hid2	h2id2	H3ID2	833	622	844	833	622	844	Individual
Storage	Not created	hj (not cleaned)	h2j	H3J (not cleaned)	833	634	845	833	634	845	Individual
Saving groups	Not created	Not included	Not included	H3LB	Not included	Not included	844	Not included	Not included	844	Individual
Designated respondent tracking	Not created	Not included	h2track	H3TRACK	Not included	25	845	Not included	25	845	Individual
Spending money	Not created	Not included	ha2f	HA3F	Not included	620	844	Not included	620	844	Individual

SHORT	Panel data	Questionnaire sections			Number of observations			Number of households			Unit of observation
		R1	R2*	R3	R1	R2*	R3	R1	R2*	R3	
Income sources	Panel AB	sab	-	SA3B	15072	-	13644	942	-	946	Income source
General outlook questions	Panel AG	sag	-	SA3G	942	-	946	942	-	946	Individual
Consumption expenditures	Panel AH	sah	-	SA3H	17841	-	19083	939	-	947	Consumption item
Household food security	Panel AI	sai	-	SA3I	942	-	946	942	-	946	Individual
Membership in groups	Panel AL	sal	-	SA3L	11304	-	10856	942	-	945	Membership group
Household Roster	Panel HA	shortq_household roster	-	SH3A	5924	-	5924	942	-	908	Household members
Assets	Panel HC	shc	-	SH3C	942	-	946	942	-	946	Household
Agricultural production	Panel HF_part1	-	-	SH3FPART2	-	-	945	-	-	945	
	Panel HF_part2	shortq	-	SH3F	942		4572	942	-	947	Agricultural
Savings	Not created	Not included	-	SH3SAVE1	Not included	-	3649	Not included	-	945	Saving type
	Not created	Not included	-	SH3SAVE2	Not included	-	938	Not included	-	938	Individual

ADDITIONAL PANELS CREATED	Panel data	Questionnaire sections	Number of observations	Number of households	Unit of observation
Consumption expenditures	Panel AHwide	ahwide (R1), sahwide (R1), A3HWIDE (R3), SA3HWIDE (R3)	3516	3516	Individual
All Purpose	All Purpose Panel Cleaned Dataset	Multiple	3649	3649	Individual

*: R2 is not included in panels

9 List of abbreviations

Soldev	Livingstonia Synod development department
RFF	Rockwool foundation research unit
TA	Traditional authority (administrative level)
ADD	Agricultural Development Division (if not otherwise specified, ADD refers to the Karonga department of ADD)
PAT	Poverty assessment tool
PPI	Progress out of poverty index

10 Bibliography

Allen, H. and Staehle, M., 2007, Programme Guide - Field Operations Manual version 3.1. VSL Associates, Stuttgarten, downloaded from vsla.net.

Ashe, J., 2002. A Symposium on Savings-Led Microfinance and the Rural Poor.Journal of Microfinance, 4 (2), 129.

Bouman, F., 1995. Rotating and accumulating savings and credit associations: A development perspective. World Development, 23 (3), 371-384.

Dorward, A. and Chirwa, E. (2011): The Malawi agricultural input subsidy programme: 2005/06 to 2008/09, *International Journal of Agricultural Sustainability*, 9(1), pp. 232-247.

Deaton, A. and Dupriez, O. (2011): Purchasing power parity exchange rates for the global poor, *American Economic Journal: Applied Economics*, 3(2), pp. 137-166.

Duflo, E., Glennerster, R. and Kremer, M. (2008): Using randomization in development economics research: A toolkit in *Handbook of Development Economics* Volume 4, pp. 3895-3962.

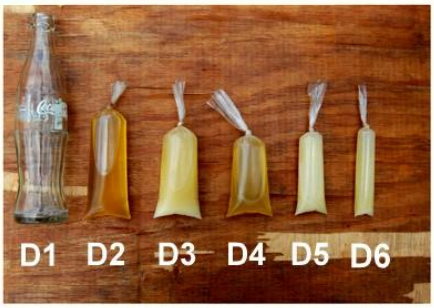
Skoufias, E., Davis, B. And Behrman, J. R. (1999): An evaluation of the selection of beneficiary households in the Education, Health and Nutrition Program (PROGRESA) of Mexico, *IFPRI working paper*

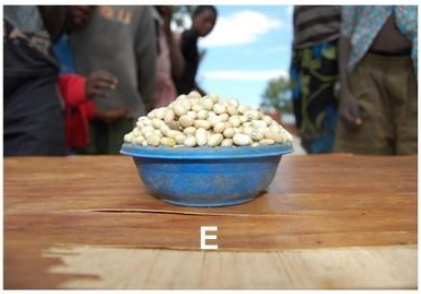
UN (2005): *Designing Household Survey Samples: Practical Guidelines*, United Nations, New York

NATURAL CANE
BROWN SUGAR
ILOVO
A1
A2
A3
A4

B1
B2

KAYSALT
KENSALT
C1
C2

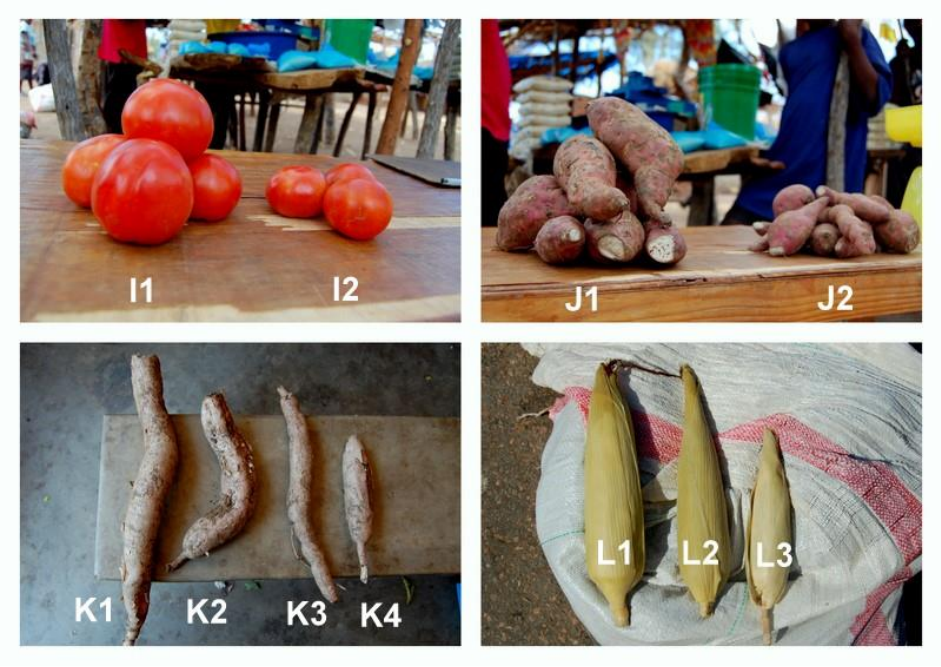

D1
D2
D3
D4
D5
D6


E

F

G

H

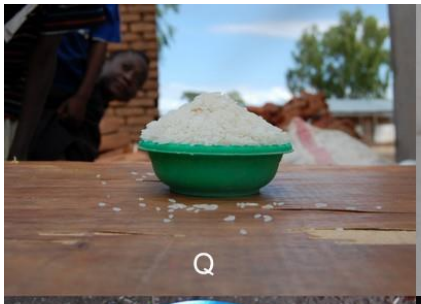

I1
I2
J1
J2
K1
K2
K3
K4
L1
L2
L3

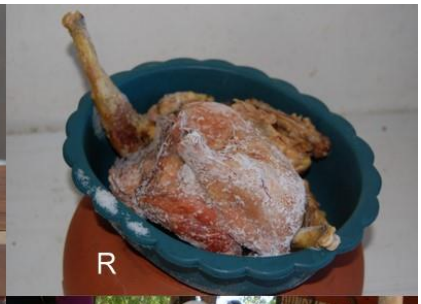
M1
M2

N

O1
O2
O3

P1
P2
P3


Q


R

Kazinga
KAZINGA
S1
S2
S3

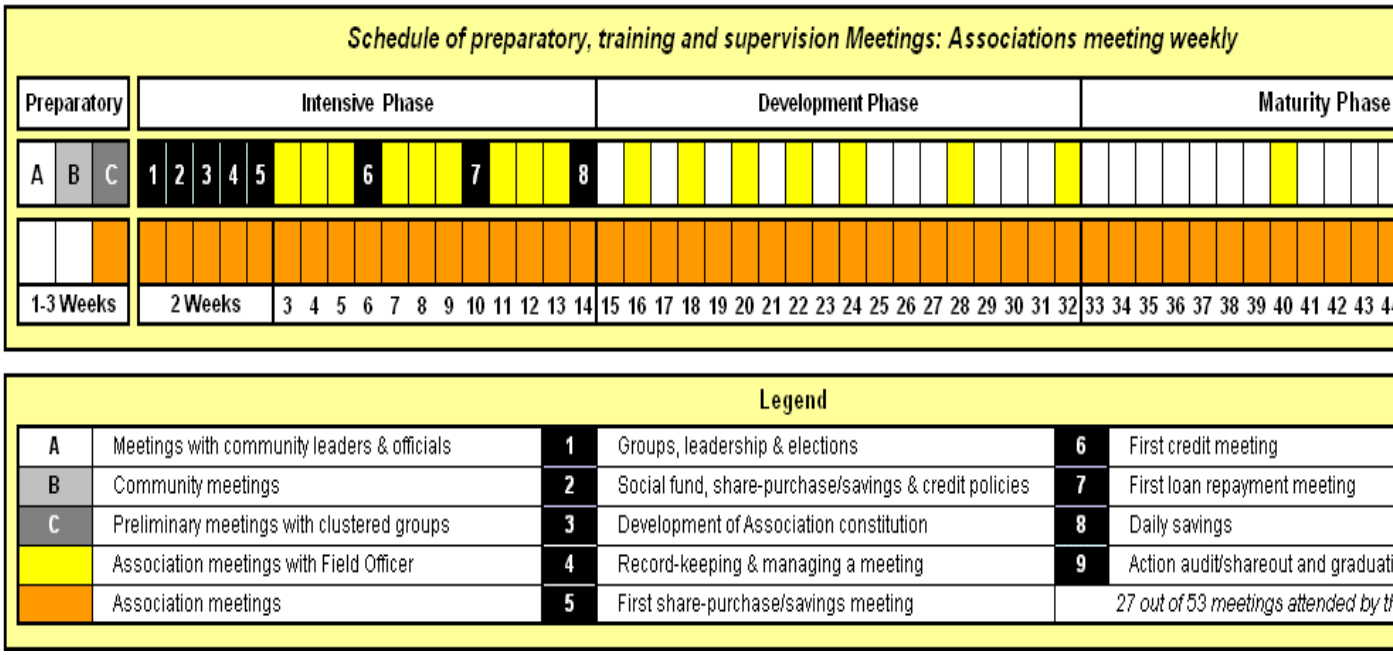
V

Z1
Z2
Z3
Z4

Z5

11 Annex 1. Schedule of training activities for VSLAs.

The schedule shows 50 group meetings. Training sessions are marked with numbers and described below the schedule. A field officer attends the meetings marked with yellow.


Schedule of preparatory, training and supervision Meetings: Associations meeting weekly
Preparatory
Intensive Phase
Development Phase
Maturity Phase
A B C
1 2 3 4 5 6 7 8
1-3 Weeks
2 Weeks
3 4 5 6 7 8 9 10 11 12 13 14
15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
33 34 35 36 37 38 39 40 41 42 43
Legend
A Meetings with community leaders & officials
B Community meetings
C Preliminary meetings with clustered groups
Association meetings with Field Officer
Association meetings
1 Groups, leadership & elections
2 Social fund, share-purchase/savings & credit policies
3 Development of Association constitution
4 Record-keeping & managing a meeting
5 First share-purchase/savings meeting
6 First credit meeting
7 First loan repayment meeting
8 Daily savings
9 Action audit/shareout and graduati
27 out of 53 meetings attended by th

Allen, H. and Staehle, M., 2007, *Programme Guide - Field Operations Manual version 3.1*. VSL Associates, Stuttgarten, downloaded from vsla.net.

Ashe, J., 2002. A Symposium on Savings-Led Microfinance and the Rural Poor. *Journal of Microfinance*, 4 (2), 129.

Bouman, F., 1995. Rotating and accumulating savings and credit associations: A development perspective. *World Development*, 23 (3), 371-384.